Fractal functional regression for classification of gene expression data by wavelets

Margarita María Rincón[1] and María Dolores Ruiz-Medina[2]

[1] University of Granada
Campus Fuente Nueva
18071 Granada, Spain
(e-mail: mmrh@correo.ugr.es)
[2] University of Granada
Campus Fuente Nueva
18071 Granada, Spain
(e-mail: mruiz@ugr.es)

Abstract. Functional classification of fractal temporal gene expression data is performed in terms of logistic regression based on the wavelet-vaguelette decomposition of the temporal gene expression curves. The fractality features of the gene expression profiles comes from the stochastic evolutionary forces acting on genomes. The noise level introduced by such forces increases local singularity, that must be removed to make robust the classification procedure. Specifically, thresholding rules are applied to the wavelet-like decomposition of the gene expression profiles to eliminate the noise. Leave-one-out cross-validation is then performed to choose the threshold minimizing the classification error.
Keywords: Fractal gene expression profiles, functional classification procedures, functional data, wavelet-vaguelette decomposition.

1 Introduction

Stochastic evolutionary forces acting on genomes induce a chaotic evolutionary structure affected by random mutations, natural selection, and genetic drifts. Serious signal to noise problems then arise, hindering the identification of similarities between anciently divergent sequences. The transferred information by homology is serious affected by this fact. Thus, the functional classification procedure must be robust against local variability induced by the noise levels in gene expression data. However, most of the functional statistical classification approaches, e.g. Functional-Principal-Component-Analysis-based classification (see Ramsay and Silverman [8], 2005) relies on the assumption of smoothness. That is, gene expression profiles are considered to be independent realizations of a smooth stochastic process (see, for example, Leng and Müller [4], 2006). In this paper we go beyond this assumption, addressing the problem of classification of functional fractal gene expression data.

It is well known that wavelet functions provide an optimal processing of chaotic structures, and, at the same time, wavelet-thresholding techniques

lead to a very effective discrimination between the structural local variability and the noise singularity/fractality (see Antoniadis and Sapatinas [2], 2003; Vidakovic [9], 1998, among others). In this paper, the temporal noising gene expression profiles are processes in terms of the wavelet-vaguelette thresholded transform. Thus, the functional gene expression classification problem is addressed taking into account the signal to noise problem. Specifically, a wavelet-thresholded-like-based functional logistic regression approach is considered to make robust the functional classification method against chaotic structure induced by stochastic evolutionary forces acting on genomes. The noise levels introduced by such forces are located in the highest levels of the *wavelet-transform-like* performed to the gene expression profiles, being then eliminated in the functional classification procedure proposed in this paper. This methodology provides low-error rate classification for the yeast cell-cycle gene expression profiles analyzed, affected by additive noise. The advantages of our approach is that fractal gene expression data with high local variability, that is, with a small biological cell signal-to-noise ratio can be suitably classified, while other smooth-based statistical approaches fail.

2 Statistical methodology

In this section, we describe some elements of wavelets theory and functional generalized linear models (FGLM). The first one as a tool for functional decomposition of biological fractal signals. In particular, in this paper, we will derive a *wavelet-like* decomposition of the gene expression profiles affected by additive biological cell noise. The second one involved in the functional statistical classification methodology proposed in this paper.

Multiresolution-like Analysis

Functional wavelet bases have been widely used in the analysis of fractal biological signals, since their provide a localized multiscale decomposition of such signals. The wavelet transform of a random biological signal $\{X(t),\ t \in \mathbb{R}\}$ leads to a sequence of correlated random wavelet coefficients. To avoid redundancy in such coefficients the random wavelet-vaguelette decomposition of a random signal is considered here (see Angulo and Ruiz-Medina [1], 1999). In this decomposition the scaling and wavelet bases are transformed to get biorthogonal bases. The transforming filter is defined from the covariance factorization of X. Specifically, representing by r_X the covariance kernel of X, given by $r_X(t,s) = E[X(t)X(s)] - E[X(t)]E[X(s)]$, the following factorization is considered

$$r_X(t,s) = \int_{\mathbb{R}} t_X(t,u) t_X(s,u) du, \tag{1}$$

in terms of kernel t_X, defining the integral operator $\mathcal{T}_X$. The transformed scaling and wavelet bases are then constructed as

$$\varphi_k(t) = \int_{\mathbb{R}} t_X(t,u)\phi_k(u)du, \quad k \in \mathbb{Z},$$
$$\gamma_{j,k}(t) = \int_{\mathbb{R}} t_X(t,u)\psi_k(u)du, \quad k \in \mathbb{Z},\ j \in \mathbb{N}. \tag{2}$$

Their dual, biorthogonal bases, are defined as

$$\varphi^k(t) = [\mathcal{T}_X^{-1}]^*(\phi_k)(t), \quad k \in \mathbb{Z},$$
$$\gamma^{j,k}(t) = [\mathcal{T}_X^{-1}]^*(\psi_k)(t), \quad k \in \mathbb{Z},\ j \in \mathbb{N}. \tag{3}$$

The projection of X in the above biorthogonal bases leads to the wavelet-vaguelette decomposition

$$X(t) = \sum_{k\in\mathbb{Z}} X^k\varphi_k(t) + X^{j,k}\gamma_{j,k}(t), \quad t \in \mathbb{R}, \tag{4}$$

where the random projections $\{X^k, \quad k \in \mathbb{Z}\}$, $\{X^{j,k},\ k \in \mathbb{Z},\ j \in \mathbb{N}\}$ are uncorrelated, that is, independent in the Gaussian case.

Functional Generalized Linear Models

Generalized Linear models (GLM)[5],[3] constitute a flexible extension of classical linear models where the response variables, $Y_1, ..., Y_N$, are independent and identically distributed (i.i.d.), with probability distribution in the exponential family. In particular, the logistic regression model is defined in terms of a set of parameters $\beta_1, ..., \beta_p$ and a set of explanatory variables $\{x_{i1}, ..., x_{ip}\}$, $i = 1, ..., N$. The monotone link function g satisfied that $E(Y_i) = \mu_i = g^{-1}(\eta_i)$, with $\eta_i = \sum x_{ij}\beta_j$. The response variables Y_i, $i = 1, ..., N$, are distributed, in this case, as a Bernoulli with mean μ_i. The link function g is chosen to be the *logit* function, given by $g(x) = log\{x/(1-x)\}$.

The maximum-likelihood estimation of the vector parameter β is derived from the maximization of the log-likelihood function $l = \sum l_i$, with l_i being the log-likelihoods associated with Y_i, for $i = 1, ..., N$. That is, vector parameter β is estimated from the log-likelihood equations, given by the identities

$$\frac{\partial l}{\beta_j} = \sum_i \frac{\partial l_i}{\beta_j} = \sum_i \frac{\partial l_i}{\partial\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial\eta_i}{\partial\beta_j} = \sum_i \frac{(y_i-\mu_i)}{\sigma^2(\mu_i)}(g^{-1})'(\eta_i)x_{ij} = 0. \tag{5}$$

The iterated weighted least squares method is usually applied in the resolution of such equations. In the gene expression data statistical classification methodology proposed in the next section, we consider the functional formulation of the above-described logistic regression model (see [6]). Specifically, the parameter β and the explanatory variables X_i, $i = 1, ..., N$, are functions satisfying $\eta_i = \alpha + \int \beta(t)X_i(t)dt$, where α is a constant, and $\beta(t)$ is square-integrable, and, for $i = 1, ..., N$, η_i defines the conditional mean $E(Y_i|X_i(t)) = \mu_i = g^{-1}(\eta_i)$, with $Var(Y_i|X_i(t)) = \sigma^2(\mu_i)$, through the *logit* function g.

3 Functional classification of gene expression curves with high external level noise

In this section, we apply functional logistic regression, in terms of the wavelet-vaguelette decomposition (4) of fractal gene expression profiles, for their functional classification. Since they are affected by biological cell noise, due to stochastic evolutionary forces, thresholding techniques are applied to remove the external noise level produced for such forces.

Application to temporal gene expression data for yeast cell cycle The temporal gene expression data (α factor synchronized) for 90 genes involved in the yeast cell cycle [7], affected by additive noise, constitute our functional data set. The gene expression is measured every 7 minutes between 0 and 119 minutes, 18 observations for each gene. It is known that 44 of these genes are related to G_1 phase regulation and 46 to the $S, S/G_2, G_2/M$ and M/G_1 phases. The 90 sample gene expression curves are assumed to be independent realizations of a mean-square integrable stochastic process $X(t)$ on $[0, S]$, $S = 119$. The local variability induced by external forces is modelled by an additive gaussian white noise process with intensity ν. Let $X_i(t_h)$ be the observation of the i-th gen at time t_h, $i = 1, \ldots, 90$, $h = 1, \ldots, 18$. A locally weighted least squares kernel estimator $\widehat{\mu}(t)$ for longitudinal data [**?**], in terms of the Epanechnikov kernel, is considered for approximate the mean function $\mu(t)$ of the random temporal gene expression function. Its covariance function $C(s,t)$ is estimated by the empirical covariance function $\widehat{C}(s,t)$, given by $\widehat{C}(t_h, t_k) = \frac{1}{90}\sum_{i=1}^{90}(X_i(t_h) - \widehat{\mu}(t_h))(X_i(t_k) - \widehat{\mu}(t_k))$ for $h \neq k$, $h, k = 1, \ldots, 18$. The empirical covariance function $\widehat{C}(s,t)$ is evaluated on a grid with $N = 64$ equally spaced points in [0,S]. Its eigenvalues (empirical eigenvalues) $\widehat{\lambda}_l$ and the corresponding eigenvectors (empirical eigenvectors) $(\widehat{\rho}_l(t_1), \ldots, \widehat{\rho}_l(t_N))$ of the resulting matrix $\widehat{C} = \widehat{C}(t_l, t_m)$, $l, m = 1, \ldots, N$, allow us to define the empirical kernel $\widehat{t}$ as a non-parametric estimator of kernel t, factorizing the covariance function $C(s,t)$ (see equation (1)) defining the temporal dependence structure of the random gene expression function. Specifically, kernel $\widehat{t}$ is defined as

$$\widehat{t}(s,t) = \sum_{m \in \mathbb{N}} \widehat{\lambda}_m^{1/2} \widehat{\rho}_m(t) \widehat{\rho}_m(s). \tag{6}$$

Formally, the kernel of the inverse $\widetilde{\mathcal{T}} = \mathcal{T}^{-1}$ of operator $\mathcal{T}$ with t kernel can then be approximated by

$$\widehat{\widetilde{t}}(s,t) = \sum_{m \in \mathbb{N}} \widehat{\lambda}_m^{-1/2} \widehat{\rho}_m(t) \widehat{\rho}_m(s). \tag{7}$$

The construction of the empirical wavelet-vaguelette function is given in terms of kernels $\widehat{t}$ and $\widehat{\widetilde{t}}$, and a given orthonormal wavelet basis. We have chosen Haar system, with *the father wavelet*, $\widetilde{\phi}(x) = I_{[0,1)}(Ex)$ and *the*

mother wavelet, $\widetilde{\psi}(x) = I_{[0,1/2)}(Ex) - I_{[1/2,1)}(Ex)$, where $E = \frac{N-1}{NS}$. So, for $j = 0, ..., log_2(N) - 1$ and $h, k = 1, ..., N$

$$\widehat{\varphi}_{0,k}(t_h) = \sum_{l=1}^{N} \widehat{t}(t_h, t_l)\widetilde{\phi}_{0,k}(t_l) \tag{8}$$

$$\widehat{\gamma}_{j,k}(t_h) = \sum_{l=1}^{N} \widehat{t}(t_h, t_l)\widetilde{\psi}_{j,k}(t_l) \tag{9}$$

In matrix form, we denote by matrix $\boldsymbol{\varphi}_0 = \{a_{hk}\}$, with $a_{hk} = \widehat{\varphi}_{0,k}(t_h)$, the product of the matrices $\widehat{\mathbf{T}} = \{b_{hk}\}$, with $b_{hk} = \widehat{t}(t_h, t_k)$, and $\boldsymbol{\Phi}_0 = \{c_{hk}\}$, with $c_{hk} = \widetilde{\phi}_{0,k-1}(t_h)$. Similarly, $\boldsymbol{\Gamma}_j = \{d_{hk}\}$, with $d_{jk} = \widehat{\gamma}_{j,k}(t_h)$, is the product of $\widehat{\mathbf{T}}$ with $\boldsymbol{\Psi}_j$, for $j = 1, \ldots, M(N)$, and $M(N) = \log_2(N)$.
Now, $\boldsymbol{\varphi}^0 = [\widehat{\mathbf{T}}^{-1}]^T \times \boldsymbol{\Phi}_0$ and $\boldsymbol{\Gamma}^j = [\widehat{\mathbf{T}}^{-1}]^T \times \boldsymbol{\Psi}_j$, for $j = 1, \ldots, M(N)$.
For X_i, the i-th gen, the following empirical coefficients are computed:

$$\widehat{\widetilde{X}}_i^{0,m} = \sum_{l=1}^{N} (X_i(t_l) - \widehat{\mu}(t_l))\widetilde{\sigma}^{0,m}(t_l),\ m = 1, \ldots, L(0) \tag{10}$$

$$\widehat{X}_i^{j,m} = \sum_{l=1}^{N} (X_i(t_l) - \widehat{\mu}(t_l))\widetilde{\gamma}^{j,m}(t_l),\ m = 1, \ldots, L(j),\ j = 1, \ldots, M(N),$$

where $L(j) = 2^j$, for $j = 1, \ldots, M(N)$, and $\widehat{\varphi}^{j,m}(t_l)$ and $\widehat{\gamma}^{j,m}(t_l)$ are the elements in the l-th row and m-th column of the matrices $\boldsymbol{\varphi}^0$ and $\boldsymbol{\Gamma}^j$, for $j = 1, \ldots, M(N)$, respectively.
Note that

$$\langle \widehat{\varphi}_{j,m}, \widetilde{\varphi}^{j,n} \rangle = \delta_{m,n} \quad \langle \widehat{\gamma}_{j,m}, \widehat{\gamma}^{j,n} \rangle = \delta_{m,n}. \tag{11}$$

Individual temporal gene expression profiles can then be approximated in terms of the following identities

$$X_i(t) = \widehat{\mu} + \sum_k \widehat{\widetilde{X}}_i^{0,k}\ \widehat{\varphi}_{0,k}(t) + \sum_j \sum_k \widehat{X}_i^{j,k}\widehat{\gamma}_{jk}(t), \quad t \in [0, S]. \tag{12}$$

Figure (1) compares the original data with their approximation by decomposition (12), for different values of ν . This decomposition will be considered in the implementation of functional logistic regression to classify the data into two groups, G_0 and G_1, using a response variable Y with Bernoulli distribution with mean μ. The response variable Y takes the value $Y = 1$ if the gene expression profile is in group G_1, or $Y = 0$ if it isn't. We define $\eta_i = \alpha + \int \beta(t)Z_i(t)dt$, $Z_i(t) = X_i(t) - \hat{\mu}_i(t)$, so $Y_i = g^{-1}(\eta_i) + e_i$, for g the *logit* function, and independent and identically distributed errors (i.i.d.) e_i, $i = 1, \ldots, 90$, with zero-mean and finite variance.

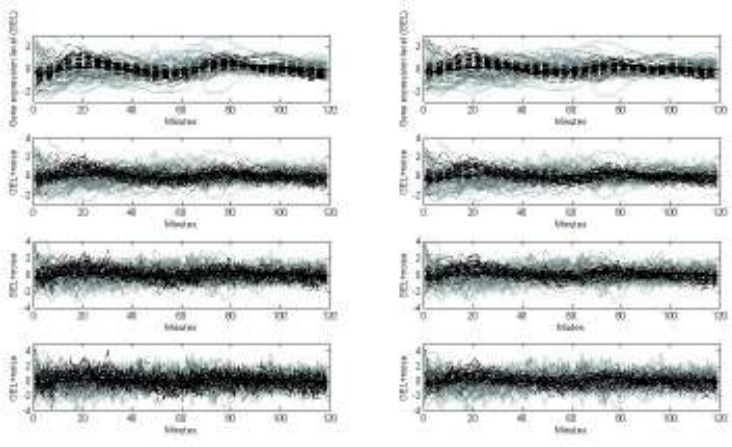

Fig. 1: Left panels: Temporal gene expression profiles of yeast cell cycle, first one: Without noise; second, third and fourth: With noise intensity $\nu = 0.35$,$\nu = 0.6$ and $\nu = 0.9$ respectively. Right panels: Reconstruction of the profiles in left panel using (12). Dashed lines: Genes expressed in G_1 phase; gray solid lines genes expressed in non-G_1 phase

Due to the square integrability of β and using (11), it is possible to write Z_i as in (12). Specifically,

$$\beta(t) = \sum_k \widetilde{\beta}_{0,k}\widehat{\gamma}^{0,k}(t) + \sum_j \sum_k \widetilde{\beta}_{j,k}\widehat{\gamma}^{jk}(t), \quad t \in [0, S], \tag{13}$$

and

$$\eta_i = \alpha + \sum_k \widehat{\widetilde{Z}}_i^{0,k} \widetilde{\beta}_{0,k} + \sum_j \sum_k \widehat{Z}_i^{j,k}\widetilde{\beta}_{jk}. \tag{14}$$

The parameters $\boldsymbol{\beta}^T = (\widetilde{\beta}_0, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_H)$, with $\widetilde{\beta}_0 = \alpha$, with $H = \sum_{j=0}^{\log_2(N)} 2^j$, can be estimated solving the score equation by iterated weighted least squares: For $i = 1, \ldots, 90$, $\sum_i (Y_i - \mu_i)(g^{-1})'(\eta_i)\widehat{\mathbf{Z}}_i^T/\sigma^2(\mu_i) = 0$, where $\widehat{\mathbf{Z}}_i^T$ is the vector of Fourier coefficients of Z_i on the vaguelette basis $\{\widehat{\varphi}^{0,k},\ k = 1, \ldots, L(0)\} \cup \{\widehat{\gamma}^{j,k},\ k = 1, \ldots, L(j),\ j = 1, \ldots, M(N)\}$

Finally, a prior probability p_0 is considered for G_0 memberships, and similarly, prior probability p_1 is considered for G_1 memberships. Thus, if $\hat{P}(Y_i = 1|X_i(t)) = g^{-1}(\hat{\alpha} + \sum_{h=1}^{H} \widehat{Z}_{i,h}\widehat{\widetilde{\beta}}_h) \geq p_1$, the i-th gen is member of G_1. Otherwise, it belongs to G_0.

Results

In order to measure the accuracy of the model, the cross-validation classification error rate (CVE) is obtained. Suppose the i-th gene is missing, the mean and the covariance function estimates, based on the other 89 genes,

and parameters β, from the reduced functional sample, are then computed. These parameters are tested to obtain the approximation η^{-i} of η, based on the sample information provided by the 89 gene expression curves, removing the i-th gene. The coefficients of the i-th gene are then computed as follows:

$$\widehat{\widehat{X}}_i^{0,m} = \sum_{l=1}^{N} (X_i(t_l) - \widehat{\mu}^{(-i)}(t_l))(\widehat{\gamma}^{0,m})^{(-i)}(t_l) \tag{15}$$

$$\widehat{X}_i^{j,m} = \sum_{l=1}^{N} (X_i(t_l) - \widehat{\mu}^{(-i)}(t_l))(\widehat{\gamma}^{j,m})^{(-i)}(t_l) \tag{16}$$

where $(\widehat{\gamma}^{0,m})^{(-i)}(t_l)$ and $(\widehat{\gamma}^{j,m})^{(-i)}(t_l)$ are computed, as in (10), but considering the empirical covariance function and the estimated mean function $\widehat{\mu}^{(-i)}$ obtained from the reduced functional sample. This procedure is repeated with every gene, if $g^{-1}(\eta^{(-i)}) \geq p_1$ the i-th gen is member of G_1, otherwise is from G_0. The CVE is defined as the quotient between the total number of genes misclassified under cross-validation, and the total number of genes. In this case, we obtain a $CVE = 0.0889$, for the gene expression profiles without external noise. In the case where such profiles are affected by the local variability of the stochastic evolutionary forces, Table 3 shows the CVE for different intensities ν of the noise, that is, for different biological cell signal-to-noise ratios, without applying thresholding (first column), and applying thresholding (second and third colums). The thresholds considered lead to the elimination of the last level $j = 5$, in second column, and the two last levels $j = 4, 5$, in third column of Table 3. The best performance is obtained in the case of $\nu = 0.35$. The rest of higher fractal cases, $\nu = 0.6$ and $\nu = 0.9$, are considerably improved eliminating the wavelet-vaguelette coefficients associated with the last *resolution-like levels*. Indeed, the results will be better when a better fitting of the threshold is obtained according to the level noise of the temporal gene expression data.

4 Final Comments

The expression of the gene is determined by different external or internal factors. Research is developed to detect and quantify gene expression levels under different scenarios (biological cells). That is, genes are expressed in different ways in different type of cells (liver cells, muscle cells, etc.). The differentiation between cells is given by the specific patterns of gene activations which in turn control the production of protein. The functional statistical classification methodology proposed in this paper is robust against the external noise factors (external dynamical forces), decreasing the biological cell signal-to-noise ratio. Thus, fractal gene expression data, associated with high local expression level variation, can be processed with the methodology proposed here, while classical functional statistical classification methods (e.g.

ν	0	$j=5$	$j=4,5$
$\nu=0$	0.0889	0.0889	0.0555
$\nu=0.35$	0.1	0.1667	0.1555
$\nu=0.6$	0.2778	0.1889	0.1
$\nu=0.9$	0.3	0.2333	0.1889
$\nu=0.35$	0.1667	0.1333	0.0667
$\nu=0.6$	0.2222	0.2	0.1555
$\nu=0.9$	0.2444	0.2222	0.1555
$\nu=0.35$	0.1333	0.1111	0.0778
$\nu=0.6$	0.2444	0.1778	0.1444
$\nu=0.9$	0.2444	0.2222	0.1555

Table 1: *CVE* eliminating coefficients associated with j for different values of the covariance ν for white noise added

Functional-Principal-Component-Analysis-based classification) are not able to process them, eliminating gene expression curves with high local variability (see, for example, Leng and Müller [4], 2006). Specifically, the classical methodologies are not able to discriminate the external noise level present in such gene expression data. This paper then addresses this problem.

Acknowledgments. This work has been supported in part by projects MTM2008-03903, of the DGI, MEC, and P06-FQM-02271 of the Andalousian CICE, Spain.

References

1.J.M. Angulo and M.D. Ruiz-Medina. Multiresolution approximation to the stochastic inverse problem. *Advances in Applied Probability. Appl. Prob*, 31:1039–1057, 1999.
2.A. Antoniadis and T. Sapatinas. Wavelet methods for continuous time prediction using hilbert-valued autoregressive processes. *Journal of Multivariate Analysis*, 87:133–158, 2003.
3.A.J. Dobson and A.G. Barnett. *An Introduction to Generalized Linear Models.* Chapman & Hall, 2008.
4.X. Leng and H.G.Muller. Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22(1):68–76, 2006.
5.P. McCullagh and J.A. Nelder. *Generalized Linear Models.* Chapman & Hall, 1989.
6.H.G. Muller and U.Stadtmuller. Generalized functional linear models. *The Annals of Statistics*, 33:774–805, 2005.
7.P.T.Spellman and et al. Comprehensive identification of cell-cycle regulated genes of the yeast Saccharomyces cerevisiae by microarray hibridization. *Mol.Biol.Cell*, 9:3273–3297, 1998.
8.J.O. Ramsay and B.W. Silverman. *Functional Data Analysis.* Springer, 2005.
9.B. Vidakovic. *Statistical Modelling by Wavelets.* Wiley Series in Probability and Statistics, 1999.